View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by OpenEdition

CORPUS

**Corpus**

**8 | 2009**
**Corpus de textes, textes en corpus**

# TEXTE ET DISCOURS. Corpus, co-texte et analyse automatique du point de vue de l'analyse de discours

**Georgeta Cislaru et Frédérique Stiri**

OpenEdition Journals

**Édition électronique**
URL : http://journals.openedition.org/corpus/1678
ISSN : 1765-3126

**Éditeur**
Bases ; corpus et langage - UMR 6039

**Édition imprimée**
Date de publication : 15 novembre 2009
Pagination : 85-104
ISSN : 1638-9808

**Référence électronique**
Georgeta Cislaru et Frédérique Stiri, « TEXTE ET DISCOURS. Corpus, co-texte et analyse automatique du point de vue de l'analyse de discours », *Corpus* [En ligne], 8 | 2009, mis en ligne le 01 juillet 2010, consulté le 01 mai 2019. URL : http://journals.openedition.org/corpus/1678

© Tous droits réservés

# TEXTE ET DISCOURS
## Corpus, co-texte et analyse automatique du point de vue de l'analyse de discours

Georgeta CISLARU
SYLED, Université Paris 3

Frédérique SITRI
SYLED, Université Paris 3 et Université Paris 10

Cet article se propose de s'interroger sur les conditions de constitution d'un corpus d'écrits de signalement d'enfant en danger du point de vue de l'analyse de discours (AD). Il s'agit de rapports éducatifs et de notes d'information visant à évaluer une situation familiale et à préconiser le cas échéant une mesure de protection pour l'enfant[1]. On a ainsi affaire à des textes qui s'inscrivent dans une pratique sociale et qui sont sous-tendus par des discours tels les textes de loi, les discours médiatiques, les guides de rédaction, les discours de la psychologie ou de la psychanalyse... La longueur des textes et la complexité des données nous ont amenées à utiliser le logiciel *Lexico 3* pour l'exploration textuelle du corpus.

Ces textes deviennent interprétables dès que le lecteur sort de leurs frontières formelles pour prendre en compte le cadre institutionnel dans lequel ils sont produits et les contraintes discursives et institutionnelles qui leur sont imposées. On remarquera d'ailleurs que les travaux de linguistique textuelle semblent être de plus en plus attentifs aux données extra-textuelles, saisies par la notion de *contexte*. Se pose alors la question de savoir comment on intègre à l'analyse ces « extérieurs » tout en s'appuyant sur les données textuelles. L'analyse de discours (AD), qui a pour objet les productions écrites ou orales, envisagées dans leur matérialité linguistique et

1 Le corpus est présenté en 1.1.

*Corpus* n°8 « Corpus de textes, textes en corpus » (2009), 85-104

dans leurs conditions de production historiques et politiques[2], offre un point de vue qui structure l'interprétation de ces textes. Nous nous proposons de montrer les incidences d'un tel point de vue sur le traitement du co(n)texte[3] dans la constitution de corpus, dans la perspective d'une exploration automatique[4] des discours.

La prise en compte du contexte associée à la possibilité matérielle de rassembler de grandes masses textuelles induit également un questionnement sur le corpus et son traitement. Or le corpus est défini comme étant nécessairement construit, préconçu. Il s'agit là d'un questionnement central en AD, car il comporte des aspects non seulement méthodologiques mais aussi théoriques : le passage de corpus clos à des corpus ouverts est lié à la prise en compte progressive de l'hétérogénéité des discours[5]. Quant à l'automatisation de la recherche[6], elle soulève des questions cruciales pour l'AD, comme celles liées à l'interprétation contextuelle des formes ; on questionnera donc ici l'impact de l'outil sur l'objet d'analyse.

## 1. La question du corpus

La constitution d'un corpus en AD soulève la question de sa contextualisation, question qui nous permettra de confronter les notions de contexte et d'interdiscours, et de nous interroger d'une part sur la « clôture » du corpus et d'autre part sur les extérieurs du texte.

2 Le discours est pris dans un réseau de discours qui l'informent et le déterminent (l'interdiscours), et qui s'inscrivent dans des formes de langues.

3 Cette graphie conjugue co-texte (environnement linguistique) et contexte (environnement extralinguistique), que l'AD envisage comme interdépendants. Voir plus loin 1.4.

4 Par exploration (ou analyse) automatique on entend ici le recours à des outils de traitement statistique (textometrie, lexicométrie…).

5 Cf. Guilhaumou et Maldidier (1979), Moirand (2004).

6 Rappelons qu'un des premiers textes de Pêcheux, théoricien de l'AD, s'intitulait Analyse automatique du discours.

### *1.1 Le corpus en AD*

L'AD favorise *grosso modo* deux démarches de constitution de corpus[7]. L'une, qu'on pourrait appeler « cadrative », part d'un cadre institutionnel et énonciatif et définit le corpus comme partie prenante de ce cadre.

> Le corpus n'y est donc pas considéré pour lui-même, mais en ce qu'il est partie prenante dans une institution reconnue qui « définit pour une aire sociale, économique, géographique ou linguistique donnée les conditions d'exercice de la fonction énonciative ». (Maingueneau 1991 : 17, citation de Foucault 1969 : 153)

L'autre, qu'on appellerait « contextualisante »[8], part de la matérialité linguistique et cherche à la situer par rapport à un genre ou par rapport à un cadre institutionnel et énonciatif. Cette démarche de contextualisation constitue un élément du processus d'interprétation des formes.

Ces approches du corpus ne sont pas antagoniques, elles se rejoignent d'ailleurs au niveau de la contextualisation par rapport à un cadre institutionnel et énonciatif. Nous verrons plus bas comment cette articulation s'est opérée dans notre étude. Il est à noter que, suite à la diffusion des travaux de Bakhtine, le cadre générique, ressenti comme plus opératoire, tend à se substituer au cadre institutionnel. En effet, le genre, produit d'une sphère sociale de l'activité humaine (Bakhtine 1984 : 265), constitue un cadre de production et d'interprétation des faits langagiers. C'est alors l'unité générique qui garantit l'unité du corpus tout en permettant la comparaison (voir corpus 1 ci-dessous).

On insistera sur le fait que le corpus ne tend pas à s'intégrer à un corpus de référence permettant de décrire la langue fonctionnelle (Rastier 2007) mais revendique une contingence contextuelle en tant qu'échantillon d'une pratique langagière.

---

7 Nous nous limitons ici aux grandes lignes.

8 On dit alors que le « corpus […] se fait l'image d'un contexte saisi sous un certain angle de vue » (Rastier et Pincemin 1999 : 84).

### *1.2 Deux corpus d'analyse*

Le corpus sur lequel nous avons travaillé est constitué d'éléments hétérogènes par leur nature aussi bien que par leur statut dans la recherche.

Un premier corpus (corpus 1 ; 9870 formes) comprend 18 rapports produits par des services sociaux, dans le cadre d'un signalement d'enfant en danger, à partir de l'observation de la situation familiale et surtout à partir d'entretiens avec l'enfant faisant l'objet du signalement, sa famille et des intervenants extérieurs (école, services sociaux, médecins, etc.). Ce corpus a donc pour spécificité de relever de l'écrit professionnel.

Les textes qui le constituent possèdent un certain nombre de caractéristiques communes. Du point de vue du dispositif communicatif, ils sont rédigés par des travailleurs sociaux (éducateurs, assistantes sociales, puéricultrices) à destination de l'instance administrative ou juridique qui a sollicité le rapport et qui possède un pouvoir de décision concernant la famille observée. Outre ce destinataire « officiel », le rapport est également lu par le chef de service qui peut demander des corrections ou des réécritures, le cas échéant par d'autres travailleurs sociaux, mais aussi par les familles qui, depuis une loi récente, ont accès à leur dossier. Du point de vue pragmatique, ces écrits ont pour fonction d'évaluer le danger ou le risque de danger couru par un enfant. Sur le plan de la composition, ils présentent une organisation en rubriques dont l'ordre et le contenu sont à peu près superposables d'un type de rapport à un autre, d'un service à un autre (ressources matérielles, histoire familiale, entretiens, analyse et conclusion).

Compte tenu de ces régularités, nous avons considéré que ces écrits pouvaient, au-delà de leur diversité, être réunis sous une même étiquette générique[9]. C'est donc cette appartenance qui les constitue en corpus.

La production de ces écrits est, d'une part, encadrée par des « guides du signalement » émanant de l'éducation nationale ou des conseils généraux et, d'autre part, régie par des textes de loi qui évoluent régulièrement. L'évolution législative est elle-

9 Sur les critères permettant de définir un genre voir Rastier (2001 : 230), von Munchöw (2001 : 115).

même corrélée à une évolution du discours « sociétal » sur la maltraitance et sur le signalement et préparée par des rapports produits par des parlementaires ou des organismes para-gouvernementaux.

Les notions de maltraitance et de signalement apparaissent ainsi comme des objets sociaux discursivement configurés, que la constitution d'un deuxième corpus comprenant des textes juridiques[10], médiatiques[11], administratifs[12] a permis de mieux cerner. Notre corpus 2 participe ainsi à la constitution d'un « savoir sur le domaine » du signalement, c'est-à-dire à la compréhension des contraintes juridiques, discursives et sociales dans lesquelles évolue le corpus 1 ; il consacre le signalement comme acte d'écriture performative, ce qui justifie la prééminence du corpus 1, qui a constitué notre objet d'analyse premier.

Il est à noter que le corpus 2 est constitué par l'analyste qui, par sa connaissance du fonctionnement des objets sociaux et les hypothèses qu'il formule, considère que tel ou tel élément doit être pris en compte. Il s'agit là d'un cadrage institutionnel et énonciatif qui renvoie à la démarche cadrative de constitution d'un corpus évoquée en 1.1.

Du point de vue de l'AD, le corpus 2 ne peut être assimilé totalement à l'interdiscours du corpus 1 : il ne permet pas de saisir clairement la façon dont le discours des écrits est traversé et constitué de discours autres dont les traces sont repérables dans des formes de langue – ce dont rend compte précisément la notion d'interdiscours :

> […] tout discours dominé est tissé de discours dominants qui lui sont intégrés, (que) les frontières discursives ne sont pas atteignables, (que) le savoir antérieur s'inscrit dans la construction d'une connaissance et se repère à

10 Le Code pénal et le Code de la famille, les arrêtés en matière d'aide et d'intervention sociale.

11 Un corpus de presse (11255 formes) constitué à partir des titres suivants : L'Express, Le Figaro, L'Humanité, Libération, Le Monde, Le Point, Le Sud-Ouest.

12 Guides de signalement et rapports administratifs (Naves-Cathala, ONED, Pécresse).

> travers des formes de langue. Autrement dit, hétérogénéité et antériorité de l'interdiscours s'inscrivent à l'intérieur même de l'intradiscours, elles n'en constituent pas le contexte. (Mazière 2005 : 58)

### *1.3 Hors-discours et interdiscours*

Dans cette perspective, on considère que la norme sociale qui sous-tend l'évaluation produite par les travailleurs sociaux se construit dans les écrits de signalement à travers des formes telles que, par exemple, la concession ou la négation. Ainsi que le montre Garnier (2008 : 82), « les énoncés concessifs, […], permettent de saisir les paramètres que constitue le discours en train de se tenir pour évaluer les personnes, leur environnement et leurs comportements ».

Dans l'exemple (1) :

(1) Le logement est fonctionnel, dispose de l'ameublement et de l'électroménager adéquat aux besoins de la famille. Nous avons cependant constaté l'aspect triste de l'appartement, dû aux couleurs sombres, mais également au fait que les volets sont souvent fermés.

le discours construit l'objet « logement » en objet d'évaluation par la mise en relation de trois propriétés – « fonctionnalité », « équipement », « aspect triste », ainsi que par l'écart que le marqueur concessif pose entre ces trois propriétés.

De même la négation construit-elle en creux une norme attendue mais pas actualisée :

(2) Lorsque nous la verrons, la fillette apparaît souriante et éveillée. Elle rentre facilement en relation. Elle se montre opposante envers sa mère, pleure et boude lorsque celle-ci n'accède pas à ses demandes. Elle va facilement vers celle-ci qui **ne la repousse pas** et peut aussi se montrer affectueuse.

L'énoncé négatif décrit bien ce qu'a observé le scripteur, mais il a aussi une dimension dialogique : il laisse entendre l'écho de l'assertion à laquelle il s'oppose (la mère repousse sa fille). On peut dire que l'énoncé négatif participe à la construction d'une « sémiologie » de la relation mère-enfant qui repose sur la présence / absence d'un comportement (alcoolisme ou non, carences affectives ou pas,

présence / absence de problèmes de comportement) et permet de ce fait l'évaluation.

Cette distinction entre contextualisation par un corpus construit par l'analyste et interdiscours saisi et configuré par des formes de langue dans le discours en train de se tenir pose la question de la clôture du corpus. Si, en effet, il convient de limiter le recours à des contextes aléatoires et subjectivement reconstitués par l'analyste dans le processus d'interprétation (cf. Mayaffre 2002), la constitution d'un corpus clos nous semble impossible. Moirand (2004 : 71 *et sq.*) souligne cet écueil en invoquant « l'impossible clôture des corpus médiatiques », où la mémoire des mots instille constamment de la verticalité dans le fil horizontal du discours. C'est sans doute la question de la distinction entre intertexte et interdiscours qui se profile ici, celle aussi de la possibilité d'une détermination univoque du sens[13].

### *1.4 Cotexte et contexte*

Si les formes linguistiques peuvent mettre en place des éléments de leur propre cadre interprétatif, en configurant l'interdiscours – dans cette optique, on dira avec Guilhaumou (2002 : 22) que le discours contient ses propres ressources interprétatives –, leur interprétation en sus peut nécessiter une co(n)textualisation.

On définira le cotexte comme l'environnement textuel immédiat d'une forme linguistique à interpréter. Le cotexte pertinent pour l'analyse peut prendre des dimensions variables ; il est constitué d'une série d'emboîtements ayant pour frontière le texte. Ainsi, dans certains cas, on élargira le cotexte jusqu'au paragraphe ou au type de rubrique, voire au paratexte.

En même temps, si on définit le texte comme une unité linguistique complexe caractérisée par la cohérence et la cohésion interne, c'est la question du contexte qui se pose. Par exemple, Charolles (1995) insiste sur la pertinence contextuelle en tant que condition et déclencheur de la cohérence d'un texte.

13 Cette position conduit à questionner la proposition de Mayaffre (2002), pour qui : « l'analyse du sens profond de cet extrait réclame ici, de manière évidente, le recours à des ressources co-textuelles que l'on ne saurait envisager de traiter différemment que les ressources intratextuelles du passage » [c'est nous qui soulignons].

De même, plus récemment, Cornish (2006, en ligne) souligne-t-il l'importance du contexte pour transformer un texte en discours : « De toute manière, le texte est souvent, sinon toujours, à la fois incomplet et indéterminé par rapport au discours qui peut en être dérivé à l'aide d'un contexte ».

On observe fréquemment, dans tous les éléments du corpus, la juxtaposition d'informations contradictoires – contradiction marquée mais pas toujours résolue par la présence de connecteurs concessifs. De manière ponctuelle, ces écrits enfreignent donc la règle de non-contradiction qui gouverne la cohérence-cohésion des textes (cf. Charolles), comme dans l'exemple ci-dessous, où les deux assertions « il n'y a pas de médecin traitant » et « il apparaît toutefois qu'un suivi médical ait bien lieu » peuvent donner lieu à des inférences contradictoires quant à la responsabilité de la mère :

(3) Concernant la prise en charge médicale, il n'y a plus de suivi P.M.I. pour les enfants contrairement à ce que Madame BOULANGER peut affirmer auprès de l'équipe du Foyer. **Il n'y a pas de médecin traitant**, Madame BOULANGER faisant plutôt appel à « SOS médecin ».
**Au vu du carnet de santé, il apparaît toutefois qu'un suivi médical ait bien lieu**, les enfants étant vus si nécessaire pour des pathologies bénignes. Les vaccins sont également à jour. La mère/Vis-à-vis des mineurs.

De fait, ces caractéristiques textuelles peuvent être rapportées aux multiples contraintes paradoxales qui pèsent sur la rédaction des rapports : pris entre la relation d'aide qu'il entretient avec la famille et le contrôle qu'il exerce de fait (Rousseau 2007), le scripteur doit produire un rapport « objectif » et « neutre » mais qui en même temps alerte sur un danger possible et argumente pour une décision d'aide éducative ou de placement ; il sait de plus que le texte qu'il rédige à destination du juge pourra être lu par la famille concernée (loi de 2002)... On le voit, c'est par la prise en compte de ce qu'on pourra appeler les « conditions de production » de ces discours que l'on peut interpréter un certain nombre de phénomènes textuels apparemment « incohérents ».

En réalité, la cohérence est dépendante de la visée pragmatique des écrits de signalement : tout ce qui est dit est

pertinent pour la compréhension dans le contexte donné (la reconstitution des objectifs, l'illocutoire textuel, Viehweger 1989 : 263), et ce qui présente un semblant d'incohérence est reconstitué de ce fait grâce à des processus inférentiels (Van de Velde 1989 : 1980).

On distinguera par conséquent deux plans de contextualisation :

1) Comme précisé plus haut, la contextualisation s'opère à différents niveaux : d'une part, les données sont contextualisées par rapport au corpus, qui constitue une émanation d'un genre[14] ; d'autre part, on contextualise par rapport au cadre institutionnel et énonciatif qui permet de préciser les conditions de production du discours[15]. Nous rappelons que nous avons affaire à un corpus écrit en situation professionnelle ; la visée pragmatique des écrits est directement liée au cadre institutionnel et à la situation de travail.

2) Dans un autre ordre d'idées, comme précisé en 1.2., le corpus 1 est contextualisé par rapport au corpus 2, ce qui permet de repérer des traces d'intertexte[16] – notamment dans les stratégies de dénomination (*mère fusionnelle* [disc. psy], *mineure enceinte et en fugue* [disc. jur.]…) – ou d'interdiscours (voir plus haut en 1.3.).

La contextualisation sur les deux plans mentionnés ci-dessus donne à voir des normes sociales qui étayent les attentes des scripteurs, des normes discursives liées à une configuration pragma-énonciative, et des normes linguistiques qui configurent les représentations des dires et des situations (catégorisation, discours rapporté…). Le contexte est ainsi appréhendé comme un cadre interprétatif complexe. Dans ce qui suit, nous poursuivrons notre réflexion sur le développement de

14 De fait, aussi bien des opérations de cotextualisation que de contextualisation peuvent être réalisées par rapport au genre. Ce qui nous occupe ici, c'est la contextualisation en tant que prise en compte des contraintes pragma-discursives qui pèsent sur un genre.

15 En effet, cette perspective modifie le point de vue sur la production linguistique et en fait non plus un texte mais un discours (Adam, 1999 : 39).

16 Et conduit donc *in fine* à une cotextualisation.

l'approche multidimensionnelle à même de traiter de cette complexité.

## 2. Du corpus au texte et retour

Compte tenu de notre choix d'utiliser un corpus en tant qu'« outil de compréhension », la constitution d'un « architexte »[17] par la mise en série de plusieurs écrits de signalement et leur traitement automatique permet de mieux cerner les particularités de ces écrits. Ainsi conçu et traité, le corpus devient un contexte interprétatif des phénomènes langagiers, tout en ouvrant sur d'autres lieux d'observations (genre, corpora contrastifs, etc.).

### *2.1 Le traitement automatique comme outil de configuration de l'objet d'analyse*

Les outils automatiques[18] permettent d'explorer des corpus volumineux, en donnant une vue en surplomb des faits lexicaux et textuels, grâce aux dictionnaires de fréquences, aux listes de segments répétés, aux fréquences relatives, aux cooccurrences et aux contextes élargis de formes présélectionnées[19].

De même, l'analyse automatique peut aider à une compréhension analytique du texte grâce à la mise en série. Pour analyser les rapports de signalement, nous avons constitué un corpus d'une vingtaine de textes que nous avons soumis aux traitements énumérés ci-dessus. D'une part, l'analyse automatique a accéléré la reconnaissance et l'apprentissage d'un genre nouveau, en dégageant des régularités d'ensemble. D'autre part, elle a favorisé la prise en compte de l'intertexte, via l'étude des segments répétés.

Cependant l'on sait que l'outil configure l'objet : ainsi, l'utilisation de l'outil automatique induit un certain nombre de préconstruits d'analyse qui reformulent la question de la co(n)textualisation, en la posant non plus en termes de cadre interprétatif *a posteriori*, mais en termes de conditions

17 Le principe d'architextualité proposé par Rastier (2001) est défini en 2.1.

18 Nous avons utilisé le logiciel Lexico3 (A. Salem, SYLED-Paris 3).

19 Nous renvoyons à Habert *et alii* (1997) pour une description détaillée des méthodes d'analyse automatique de corpus.

interprétatives *a priori*. Les opérations interprétatives préalables (cf. Viprey 2006) ciblent des objets d'analyse prédéterminés et souvent constitués en listes closes. Car le balisage des corpus, tout comme le choix des formes linguistiques dont on observe les cooccurrences, relèvent déjà d'une opération interprétative[20], comme le montrent notamment les difficultés de repérage de certaines formes de discours rapporté : c'est le cas pour certaines formes marquées mais dont l'interprétation n'est pas univoque (guillemets par exemple) ou pour les formes non marquées (voir Sitri, 2008, p. 85 *sq*.). Dans ces cas de figure, il faut faire appel à une une interprétation préalable, c'est-à-dire à une analyse qualitative.

Le balisage des formes syntaxiques, la lemmatisation, la racinisation et le choix des formes lexicales à soumettre à l'analyse textométrique préconfigurent les données. Or, non seulement un nombre considérable de formes pose des problèmes de repérage, mais de plus cette manière d'approcher un texte ou un corpus ne prend pas en compte le fonctionnement en réseau des formes linguistiques, notamment lorsqu'il s'agit d'étudier la cohérence / cohésion des textes. Car, si l'identification quantitative de certaines formes linguistiques est pleinement justifiée pour l'étude de ces mêmes formes, elle ne l'est plus pour l'analyse du texte comme unité linguistique :

> Statistical linguistic analysis of this kind [counting verbs, nouns, measuring the lengh and complexity of sentences] ignores the functions of texts in communication and the pursuit of human goals. (De Beaugrande & Dressler 1981: 183)

De ce fait, analyse qualitative et analyse quantitative sont complémentaires. Deux exemples permettront d'illustrer cette complémentarité des analyses ainsi que les problèmes de balisage et d'interprétation qu'elles posent : l'analyse des

20 « […] il est non moins évident que toute lemmatisation (et plus largement toute annotation liée à une identification linguistique) est une opération interprétative, qui risque de se révéler intrusive si elle n'est pas maîtrisée comme telle. » (Viprey 2006)

formes de représentation du discours autre (RDA) et de l'expression des émotions[21].

La fréquence et l'intérêt des formes de RDA résultent du fait que l'évaluation de la situation s'appuie essentiellement sur des entretiens avec l'enfant et la famille. L'étude textométrique de la répartition des formes de RDA les plus courantes, discours direct (DD) et discours indirect (DI), fait apparaître leur fréquence relative en fonction des catégories de rapports, des rubriques (rubriques « entretiens » *vs* les autres), éventuellement des services (certains services utilisent plus le DD) et des locuteurs dont les dires sont représentés (usage fréquent du DD pour la famille et les enfants *vs* DI pour les intervenants extérieurs). Cependant si l'automatisation permet d'objectiver les quantifications, l'interprétation des effets produits par les formes nécessite assez rapidement une prise en compte de leur contexte d'occurrence (cf. Cislaru et Sitri 2008).

Par ailleurs le caractère interprétatif du repérage même de la plupart des formes de RDA pose un problème crucial du point de l'automatisation : ainsi en est-il du guillemetage, interprétable comme emprunt ou comme mise à distance, des verbes introducteurs de discours indirect, dont l'extension varie selon les genres et les contextes, sans parler de formes « non marquées » que des indices génériques et contextuels amèneront à interpréter comme constituant la suite d'un discours indirect, d'une modalisation en discours second, ou encore un discours indirect libre.

Ainsi, si le recours à l'analyse automatique permet d'objectiver des données et fournit les conditions de circulation dans un corpus, l'outil s'avère problématique quand on a affaire à des formes dont le repérage s'appuie sur des indices contextuels.

En partant du constat que le danger et le risque de danger occupent une place centrale dans les rapports de signalement, nous avons regardé de près la représentation (expression et description) de la peur dans ces textes. Dès le départ s'est posé un problème méthodologique : qu'allait-on

21 Pour un traitement plus détaillé de ces deux aspects, nous renvoyons à nos articles respectifs dans *Les Carnets du Cediscor* 10.

considérer comme des représentations de la peur ? Le silence serait-t-il représenté ? Et les manifestations psychophysiques ? Nous avons décidé de prendre pour point de départ le lexique de la peur, en tant que formes observables et quantifiables, tout en prenant la mesure des insuffisances de ce choix.

En effet, nous n'avons pu échapper aux opérations d'interprétation prédéterminantes. Ainsi, les recherches lexicométriques ont dû être partiellement faites « à la main », afin de couvrir l'ensemble des lexèmes concernés, à partir d'une liste que nous avons établie en nous appuyant sur des études lexico-sémantiques de la peur.

Deux observations s'imposent ici : d'une part, en partant de l'étude lexico-sémantique comme préalable à l'analyse lexicométrique, on fait plus immédiatement attention aux formes absentes : aucune occurrence de *horrible, terrible, affolé,* par exemple ; d'autre part, on est plus réticent à la lemmatisation et à la racinisation (les formes *inquiétant* et *inquiétude(s)* relevant, du point de vue énonciatif, de deux pôles différents, la première renvoyant au travailleur social et la deuxième aux membres de la famille).

Les exemples fournis montrent que, au-delà de la dimension subjective de construction du corpus, le traitement de ce dernier n'est pas exempt d'*a priori* subjectifs et contingents, bien au contraire.

Ainsi, le principe du hasard, appelé à contourner la subjectivité, peut intervenir directement dans l'interprétation. L'approche analytique décrite ici soulève une autre question, qui concerne le principe de l'architextualité (Rastier 2001 : 92). Tel que formulé par Rastier, ce principe[22] présume que « tout texte placé dans un corpus en reçoit les déterminations sémantiques, et modifie potentiellement le sens de chacun des textes qui le composent »[23]. Cette détermination sémantique

22 Qui, sous certains aspects, pourrait être rapproché de la notion d'hypertexte (Legallois 2006).

23 Contrairement à ce que soutient Mayaffre (2002 : 16) : « Le corpus est un objet heuristique. C'est une construction arbitraire, une composition relative qui n'a de sens, de valeur et de pertinence qu'au regard des questions qu'on va lui poser, des réponses que l'on cherche, des résultats que l'on va trouver. »

réciproque peut être observée aussi bien au niveau interprétatif qu'au niveau de la sélection des objets de recherche.

Ainsi, l'expression de la peur par les travailleurs sociaux dans le corpus 1 est contextualisée par rapport au genre *écrit social* en tant que caractérisé par une visée argumentative de « faire intervenir ». Or, la contextualisation par rapport à des éléments du corpus 2, tels le Code pénal ou le discours de presse, fait émerger d'autres interprétations. D'une part, l'inquiétude du scripteur face à la situation évaluée pourrait être en lien avec ses obligations : devoir de signalement inscrit dans la loi, responsabilité pénale en cas de non-intervention et minimisation du risque de danger (article 434-3 du Code pénal)[24]. D'autre part, la posture si proche d'une médiatisation qui n'est pas tendre avec le champ social (généralement, le signalement constitue un objet médiatique – dans la rubrique « faits divers » – lorsqu'il est trop tard pour sauver l'enfant) et qui confond facilement signalement et délation pourrait déclencher l'expression émotionnelle. Enfin, l'enchaînement de l'expression émotionnelle *enfant* / *famille* → *travailleur social* (voir plus bas 2.2.) conduit à s'interroger sur l'empathie discursive et renvoie éventuellement à des discours psychologiques.

### *2.2 Texte, genre, corpora*

La détermination sémantique du texte due à la manière de constituer le corpus est propre aussi bien à l'analyse qualitative qu'à l'analyse quantitative, mais c'est cette dernière qui permet de mieux s'en rendre compte. Ainsi, c'est l'analyse automatique – et donc, la mise en série des textes – qui nous a permis de mettre en évidence la fréquence des termes de la peur et la diversité des structures concernées. Cela nous a conduites à une analyse qualitative des structures syntaxiques d'expression / description de la peur, en distinguant les structures réflexives (*X a peur de Y*) et les structures allocentrées (*Z a peur pour X à cause de Y*) (voir Cislaru 2008). L'analyse des séquences d'apparition de ces structures à permis de dégager le pouvoir

24 Cette hypothèse a été confirmée par les travailleurs sociaux eux-mêmes lors de l'Assemblée générale de l'association « Echanger autrement » de Caen, Calvados.

cohésif de leur enchaînement au niveau du texte : on a pu ainsi observer un passage effectif des structures syntaxiques réflexives, où l'enfant ou un membre de la famille dit « *J'ai peur→L. a peur / dit avoir peur/nous parle de ses angoisses* », aux structures allocentrées, où le travailleurs social confirme « *Nous ne pouvons qu'être inquiets* », pour aboutir, dans la conclusion du rapport, à la formulation « *La situation est inquiétante* », qui correspond déjà à l'évaluation d'un risque de danger et argumente en faveur d'une mesure prise par le juge. Au niveau de l'analyse du texte en tant qu'unité linguistique, s'impose donc le constat suivant : que les émotions soient représentées dans le discours cité (description des émotions de l'enfant ou de la famille) ou dans le discours citant (expression des émotions par le travailleur social), elles ont la même orientation interprétative du point de vue de la cohérence du texte : il faut agir. Cette observation confirme le rôle cohésif du discours rapporté et de la subjectivité émotionnelle. Mais quelle conclusion tirer quant à l'influence de l'exploration textuelle sur l'interprétation, au-delà de la prédétermination ?

De fait, la visée émotionnelle de ces textes s'inscrit bien dans une perspective générique, ce qui implique un retour au corpus (corpus 1) en tant qu'émanation d'un genre discursif. Or les spécificités d'un genre discursif ne sont réellement accessibles que par la comparaison des genres. Par exemple, l'analyse contrastive des indices interprétatifs émotionnels dans un corpus d'écrits de signalement et dans un corpus de presse constitué d'articles traitant de risques et de dangers divers (11 septembre, AZF, risques alimentaires…) nous a permis de formuler des hypothèses plus fines quant à la visée pragmatique des textes relevant des deux corpus (voir Cislaru 2009). Ainsi, le lecteur des écrits de signalement est orienté vers une interprétation des émotions exprimées par le scripteur-travailleur social ou prises en charge par celui-ci. Cela peut conduire à une *offre* de protection face au danger ou risque de danger encouru par l'enfant, offre virtuelle dans le cas du lecteur extérieur à l'institution du signalement, offre bien réelle de la part du juge pour enfants qui est le destinataire de ces écrits. Le lecteur du discours de presse est quant à lui orienté vers une interprétation des émotions perspectivisées des

témoins, renforcées par un usage généralisant de *on*, ce qui peut conduire à une *demande* – là encore, réelle ou virtuelle – de protection face au danger.

On voit ainsi le corpus s'ouvrir vers des extérieurs qui ne lui sont attachés que par des hypothèses interprétatives et conduire à la constitution de corpora contrastifs. Dans cette optique, les écrits de signalement peuvent s'intégrer à une série de genres « thématiques » assimilant discours sécuritaires, discours protectionnistes, discours de défense, etc. ; le thème correspondant dans ce cadre à un « discours sur ».

Cette boucle analytique rejoint l'herméneutique philologique telle que définie par Rastier (2001 : 83) :

> Dans la perspective caractérisante de l'herméneutique philologique, un texte isolé n'est guerre interprétable, et la *collection critique* des textes s'impose. Elle n'a pas pour but d'accumuler des *data*, comme le croient certains décideurs, mais de rendre les textes lisibles, tant il est vrai que textualité et intertextualité sont interdépendantes. Tous texte, si singulier soit-il, ne se laisse comprendre qu'au sein de la multiplicité des textes du même genre et du même discours […].

## 3. Conclusion

Le corpus contribue directement à la construction de l'objet linguistique. De fait, ce que l'on observe, ce sont des ajustements successifs entre l'objet et le lieu de son observation.

On insistera aussi sur le fait que l'analyse automatique, de par sa vision en surplomb, contribue à transformer le corpus en contexte interprétatif, grâce aux régularités génériques qu'elle met en évidence et qu'il devient ensuite possible d'appliquer à l'analyse d'un texte afin d'en évaluer la cohérence et de dégager les principes sur lesquels cette dernière prend appui.

Cependant, si le corpus peut se constituer en outil d'interprétation d'un texte, il n'épuise pas pour autant le contexte, qui reste une donnée ouverte. De même, le fait que le

discours construise son propre hors-discours à travers une série de formes linguistiques garantit un garde-fou assez fiable contre une contextualisation aléatoire[25] du corpus.

Dans la continuité de ces observations, nous formulons quelques remarques pour une approche multidimensionnelle de la co(n)textualité, prenant appui sur les conditions de constitution d'un corpus non linéaire[26] et sur les enjeux et questionnements de l'analyse automatique. Premièrement, dans le cadre de cette approche, le corpus n'annule pas le texte : l'analyse sur corpus peut intégrer une étude des textes en tant qu'unités linguistiques d'une part, en tant que représentants d'un genre d'autre part[27]. En effet, le texte cotextualise les formes linguistiques en étant lui-même co(n)textualisé par les genres. Deuxièmement – et en lien direct avec le premier point –, on distingue plusieurs niveaux de co(n)textualisation qui sont appelés à faire le lien entre texte, corpora et genre. Le corpus étire et dilue partiellement les frontières du texte en proposant de l'identifier non plus par rapport à ses intérieurs (cohésion, paragraphes, etc.) mais par rapport à ses extérieurs. Mais, parce que l'analyse de discours a depuis toujours orienté le texte vers ses extérieurs, elle réussit à le préserver en tant qu'unité d'analyse sans le « noyer » dans le corpus. L'analyse automatique permet de configurer ces liens allant du corpus au texte. Bien évidemment, il ne faut pas perdre de vue le fait que, aussi bien le corpus que l'objet d'analyse, ainsi que les différents niveaux de co(n)textualisation susmentionnés sont construits par l'analyse à partir d'une série de considérations théoriques ; de même, texte, corpus et outils sont des artefacts

25 Même si, à titre expérimental, la contextualisation aléatoire reste une possibilité analytique.

26 C'est-à-dire qui intègre la dimension verticale de l'interdiscours.

27 Pour le dire en d'autres termes, l'objectif est de ne pas « manquer le texte » : « Manquer le texte pour l'analyse du discours, c'est prendre en considération les conditions de production des textes et négliger les productions elles-mêmes. Manquer le texte, pour l'herméneutique traditionnelle, c'est prétendre toucher l'âme des textes en négligeant leur chair. Manquer le texte pour la rhétorique, par exemple, c'est s'éblouir sur quelques fleurs de langage ou figure de style remarquables, lorsque la matérialité du texte, dans son ensemble, participe de l'éloquence du discours. » (Mayaffre 2007 : 23).

qui, de par ce statut commun, font de l'interprétation une construction de l'analyste. Mais l'interprétation, en étant construite par l'outil et par le lieu d'observation, n'est pas détachée du texte : elle est partie prenante du texte et du discours, intégrée au même processus de construction que ces derniers.

**Références bibliographiques**

Adam J.-M. (1999). *Linguistique textuelle. Des genres de discours aux textes*. Paris : Nathan.

Adam J.-M. (2006). « Autour du concept de texte. Pour un dialogue des disciplines de l'analyse des données textuelles », *Lexicometrica* – actes JADT'2006, Viprey J.M. (éd.), consultables en ligne à l'adresse : http://www.cavi.univ-paris3.fr/lexicometrica/jadt/index.htm

Authier-Revuz J. (2001). « Le discours rapporté », *in* Thomassone R. (éd.) *Une langue, le français*. Paris : Hachette, 192-201.

Bakhtine M. (1984 [1934]). *Esthétique de la création verbale*. Paris : Gallimard.

Beaugrande R. de, Dressler W. (1981). *Introduction to Text Linguistics*. London / New York : Longman.

Charolles M. (1989). « Coherence as a Principle in the Regulation of Discursive Production », *in* Heydrich W., Neubauer, F., Petöfi, J. S., Sözer, E. (eds.), *Connexity and Coherence. Analysis of Text and Discourse*. Berlin : de Gruyter, 3-15.

Charolles M. (1995). « Cohérence, pertinence et intégration conceptuelle », *in* Lane Ph. (éd.), *Des discours aux textes : modèles et analyses*. Rouen : Publications des Universités de Rouen et du Havre, 39-74.

Cislaru G. (2008). « L'intersubjectivation comme source de sens : expression et description de la peur dans les écrits de signalement », *Les Carnets du Cediscor* 10 : 117-136.

Cislaru G. (2009, sous presse). « Expression de la peur et interprétations sémantiques en contexte », Mémoires de la

Société Néophilologique. Helsinki : Modern Language Society.

Cislaru G., Pugnière F. et Sitri F. (dir.) (2008). *Les Carnets du Cediscor* 10 (« Analyse de discours et demande sociale : le cas des écrits de signalement »). Paris : PSN.

Cislaru G., Sitri F. (2008). « La représentation du discours autre dans des signalements d'enfants en danger : une parole interprétée ? », *Circulation des discours et liens sociaux. Le discours rapporté comme pratique sociale* (5-7 octobre 2006, Université Laval). Québec : Editions Nota Bene.

Cornish F. (2006). « Relations de cohérence en discours : critères de reconnaissance, caractérisation et articulation cohésion-cohérence », *Corela*, Numéros spéciaux, Organisation des textes et cohérence des discours. Disponible en ligne à l'URL : http://edel.univ-poitiers.fr/corela/document.php?id=1280.

Garnier S. (2008). « L'évaluation dans les rapports de signalement », *Les Carnets du Cediscor* 10 : 79-91.

Guilhaumou J. (2002). « Le corpus en analyse de discours : perspective historique », *Corpus* 1 : 21-49.

Guilhaumou J., Maldidier D. (1979). « Courte critique pour une longue histoire. L'analyse du discours ou les (mal)heures de l'analogie », *Dialectiques* 26 : 7-23.

Legallois D. (2006). « L'hypertextualité et virtualité comme modes de la construction des discours et des connaissances », *Pratiques* 129-130 : 139-156.

Habert B., Nazarenko A., Salem A. (1997). *Les linguistiques de corpus*. Paris : Armand Colin.

Maingueneau D. (1991). *L'Analyse du discours : introduction aux lectures de l'archive*. Paris : Hachette Supérieur.

Maldidier D. (1990). *L'inquiétude du discours*, textes de M. Pêcheux. Paris : Edition des cendres.

Mayaffre D. (2002). « Les corpus réflexifs : entre architextualité et hypertextualité », *Corpus* 1 : 51-69. URL : http://corpus.revues.org/document11.html.

Mayaffre D. (2007). « Philologie et/ou herméneutique numérique : nouveaux concepts pour de nouvelles pratiques ? », *in* Rastier F. et Ballabriga M. (dir.), *Corpus en Lettres et Sciences sociales*. Toulouse : PU de Toulouse Le Mirail, 15-25.

Mazière F. (2005). *L'Analyse du discours. Histoire et pratiques.* Paris : PUF.

Moirand S. (2004). « L'impossible clôture des corpus médiatiques. La mise au jour des observables entre catégorisation et contextualisation », *TRANEL* 40 : 71-92.

Munchöw P. von (2001). *Contribution à la construction d'une linguistique de discours comparative : entrées dans le journal télévisé français et allemand*, Thèse pour le doctorat, Université Paris 3 Sorbonne nouvelle.

Rastier F. (2001). *Arts et sciences du texte*. Paris : PUF.

Rastier F. (2007). « Le corpus en questions », *in* Rastier F. et Ballabriga M. (dir.), *Corpus en Lettres et Sciences sociales*. Toulouse : PU de Toulouse Le Mirail, viii-xiii.

Rastier F., Pincemin B. (1999). « Des genres à l'intertexte », *Cahiers de praxématique* 33 : 83-111.

Rousseau P. (2007). *Pratique des écrits et écriture des pratiques*. Paris : L'Harmattan.

Sitri F. (2008). « Observer et évaluer dans les rapports éducatifs : de la représentation d'un dire singulier à la description d'une situation », *Les Carnets du Cediscor* 10 : 95-116.

Van de Velde R. G. (1989). « Man, Verbal Text, Inferencing, and Coherence », *in* Heydrich W., Neubauer F., Petöfi J.-S., Sözer E. (eds.) *Connexity and Coherence. Analysis of Text and Discourse*. Berlin : de Gruyter, 174-217.

Viehweger D. (1989). « Coherence – Interaction of Modules », *in* Heydrich W., Neubauer F., Petöfi J.-S., Sözer E. (eds.), *Connexity and Coherence. Analysis of Text and Discourse*. Berlin : de Gruyter, 256-274.

Viprey J.-M. (2006). « Quelle place pour les sciences des textes dans l'Analyse de Discours ? », *Semen* 21 : 167-182.